Evaluating the Quality of RDF Data Sets on Common Vocabularies in the Social, Behavioral, and Economic Sciences

Thomas Hartmann[1], Benjamin Zapilko[1], Joachim Wackerow[1], and Kai Eckert[2]

[1] GESIS – Leibniz Institute for the Social Sciences, Germany
{firstname.lastname}@gesis.org,
[2] University of Mannheim, Germany
kai@informatik.uni-mannheim.de

Abstract. From 2012 to 2015 together with other Linked Data community members and experts from the social, behavioral, and economic sciences (*SBE*), we developed diverse vocabularies to represent SBE metadata and tabular data in RDF. The *DDI-RDF Discovery Vocabulary (DDI-RDF)* is designed to support the dissemination, management, and reuse of unit-record data, i.e., data about individuals, households, and businesses, collected in form of responses to studies and archived for research purposes. The *RDF Data Cube Vocabulary (QB)* is a W3C recommendation for expressing *data cubes*, i.e. multi-dimensional aggregate data and its metadata. *Physical Data Description (PHDD)* is a vocabulary to model data in rectangular format, i.e., tabular data. The data could either be represented in records with character-separated values (CSV) or fixed length. The *Simple Knowledge Organization System (SKOS)* is a vocabulary to build knowledge organization systems such as thesauri, classification schemes, and taxonomies. XKOS is a SKOS extension to describe formal statistical classifications.
To ensure high quality of and trust in both metadata and data, their representation in RDF must satisfy certain criteria - specified in terms of RDF constraints. In this paper, we evaluate the data quality of 15,694 data sets (4.26 billion triples) of research data for the social, behavioral, and economic sciences obtained from 33 SPARQL endpoints. We checked 115 constraints on three different and representative SBE vocabularies (DDI-RDF, QB, and SKOS) by means of the *RDF Validator*, a validation environment which is available at http://purl.org/net/rdfval-demo.

Keywords: RDF Validation, RDF Constraints, DDI-RDF Discovery Vocabulary, RDF Data Cube Vocabulary, Thesauri, SKOS, Linked Data, Semantic Web

1 Introduction

For constraint formulation and RDF data validation, several languages exist or are currently developed. *Shape Expressions (ShEx)*, *Resource Shapes (ReSh)*, *Description Set Profiles (DSP)*, *OWL 2*, the *SPARQL Inferencing Notation (SPIN)*,

and *SPARQL* are the six most promising and widely used constraint languages. OWL 2 is used as a constraint language under the closed-world and unique name assumptions. The W3C currently develops *SHACL*, an RDF vocabulary for describing RDF graph structures. With its direct support of validation via SPARQL, SPIN is very popular and certainly plays an important role for future developments in this field. It is particularly interesting as a means to validate arbitrary constraint languages by mapping them to SPARQL [4]. Yet, there is no clear favorite and none of the languages is able to meet all requirements raised by data practitioners. Further research and development therefore is needed.

In 2013, the W3C organized the RDF Validation Workshop,[3] where experts from industry, government, and academia discussed first use cases for constraint formulation and RDF data validation. In 2014, two working groups on RDF validation have been established to develop a language to express constraints on RDF data: the *W3C RDF Data Shapes Working Group*[4] (33 participants of 19 organizations) and the *DCMI RDF Application Profiles Task Group*[5] (29 people of 22 organizations) which among others bundles the requirements of data institutions of the cultural heritage sector and the *social, behavioral, and economic (SBE)* sciences and represents them in the W3C group.

Within the DCMI task group, a collaboratively curated database of RDF validation requirements[6] has been created which contains the findings of the working groups based on various case studies provided by data institutions [3]. It is publicly available and open for further contributions. The database connects requirements to use cases, case studies, and implementations and forms the basis of this paper. We distinguish 81 requirements to formulate constraints on RDF data; each of them corresponding to a constraint type.

We collected constraints for commonly used vocabularies in the SBE domain, either from the vocabularies themselves or from domain and data experts, in order to gain a better understanding about the role of certain requirements for data quality and to direct the further development of constraint languages. All in all, this lead to 115 constraints we implemented on three vocabularies. We let the experts classify the constraints according to the severity of their violation.

As we do not want to base our conclusions on the evaluation of vocabularies and constraint definitions alone, we conducted a large-scale experiment. For all these implemented 115 constraints, we evaluated the data quality of 15,694 data sets (4.26 billion triples) of SBE research data on three common vocabularies in SBE sciences (DDI-RDF, QB, SKOS) obtained from 33 SPARQL endpoints.

2 Common Vocabularies in SBE Sciences

We took all well-established and newly developed SBE vocabularies into account and defined constraints for three vocabularies commonly used in the SBE sciences

[3] http://www.w3.org/2012/12/rdf-val/
[4] http://www.w3.org/2014/rds/charter
[5] http://wiki.dublincore.org/index.php/RDF-Application-Profiles
[6] Online available at: http://purl.org/net/rdf-validation

which are briefly introduced in the following. We analyzed actual data according to constraint violations, as for these vocabularies large data sets are already published.

SBE sciences require high-quality data for their empirical research. For more than a decade, members of the SBE community have been developing and using a metadata standard, composed of almost twelve hundred metadata fields, known as the *Data Documentation Initiative (DDI)*, [7] an XML format to disseminate, manage, and reuse data collected and archived for research [8]. In XML, the definition of schemas containing constraints and the validation of data according to these constraints is commonly used to ensure a certain level of data quality. With the rise of the Web of Data, data professionals and institutions are very interested in having their data be discovered and used by publishing their data directly in RDF or at least publish accurate metadata about their data to facilitate data integration. Therefore, not only established vocabularies like SKOS are used; recently, members of the SBE and Linked Data community developed with the *DDI-RDF Discovery Vocabulary (DDI-RDF)*[8] a means to expose *DDI* metadata as Linked Data.

The data most often used in research within SBE sciences is *unit-record data*, i.e., data collected about individuals, businesses, and households, in form of responses to studies or taken from administrative registers such as hospital records, registers of births and deaths. A *study* represents the process by which a data set was generated or collected. The range of unit-record data is very broad - including census, education, health data and business, social, and labor force surveys. This type of research data is held within data archives or data libraries after it has been collected, so that it may be reused by future researchers. By its nature, unit-record data is highly confidential and access is often only permitted for qualified researchers who must apply for access. Researchers typically represent their results as aggregated data in form of multi-dimensional tables with only a few columns: so-called *variables* such as *sex* or *age*. Aggregated data, which answers particular research questions, is derived from unit-record data by statistics on groups or aggregates such as frequencies and arithmetic means. The purpose of publicly available aggregated data is to get a first overview and to gain an interest in further analyses on the underlying unit-record data. For more detailed analyses, researchers refer to unit-record data including additional variables needed to answer subsequent research questions.

Formal childcare is an example of an aggregated variable which captures the measured availability of childcare services in percent over the population in European Union member states by the dimensions *year*, *duration*, *age* of the child, and *country*. Variables are constructed out of values (of one or multiple datatypes) and/or code lists. The variable *age*, e.g., may be represented by values of the datatype *xsd:nonNegativeInteger* or by a code list of age clusters (e.g., '0 to 10' and '11 to 20'). The *RDF QB Vocabulary (QB)*[9] is a W3C rec-

[7] http://www.ddialliance.org/Specification/
[8] http://rdf-vocabulary.ddialliance.org/discovery.html
[9] http://www.w3.org/TR/vocab-data-cube/

ommendation for representing *QBs*, i.e., multi-dimensional aggregated data, in RDF [6]. A *qb:DataStructureDefinition* contains metadata of the data collection. The variable *formal childcare* is modeled as *qb:measure*, since it stands for what has been measured in the data collection. *Year*, *duration*, *age*, and *country* are *qb:dimensions*. Data values, i.e., the availability of childcare services in percent over the population, are collected in a *qb:DataSet*. Each data value is represented inside a *qb:Observation* which contains values for each dimension.

For more detailed analyses we refer to the underlying unit-record data. The aggregated variable *formal childcare* is calculated on the basis of six unit-record variables (i.a., *Education at pre-school*) for which detailed metadata is given (i.a., code lists) enabling researchers to replicate the results shown in aggregated data tables. *DDI-RDF* is used to represent metadata on unit-record data in RDF. The study (*disco:Study*) for which the unit-record data has been collected contains eight data sets (*disco:LogicalDataSet*) including variables (*disco:Variable*) like the six ones needed to calculate the variable *formal childcare*.

The *Simple Knowledge Organization System (SKOS)* is reused to a large extend to build SBE vocabularies. The codes of the variable *Education at pre-school* are modeled as *skos:Concepts* and a *skos:OrderedCollection* organizes them in a particular order within a *skos:memberList*. A variable may be associated with a theoretical concept (*skos:Concept*) and *skos:narrower* builds the hierarchy of theoretical concepts within a *skos:ConceptScheme* of a study. The variable *Education at pre-school* is assigned to the theoretical concept *Child Care* which is a narrower concept of the top concept *Education*. Controlled vocabularies (*skos:ConceptScheme*), serving as extension and reuse mechanism, organize types (*skos:Concept*) of descriptive statistics (*disco:SummaryStatistics*) like minimum, maximum, and arithmetic mean.

3 Classification of Constraint Types and Constraints

To gain better insights into the role that certain types of constraints play for the quality of RDF data, we use two simple classifications: on the one hand, we classify RDF constraint types whether they are expressible by different types of constraint languages and on the other hand, we classify constraints formulated for a given vocabulary according to the perceived severity of their violation.

Within the working groups, we identified by today 81 requirements to formulate RDF constraints (e.g., *R-75: minimum qualified cardinality restrictions*); each of them corresponding to an RDF constraint type.[10] Within a technical report, we explain each requirement/constraint type in detail and give examples for each expressed by different constraint languages [5]. We provide mappings to representations in Description Logics (DL) [2] to logically underpin each requirement and to determine which DL constructs are needed to express each constraint type. For the three vocabularies, several SBE domain experts determined the default severity level of the 115 concrete constraints, which we pub-

[10] Constraint types and constraints are uniquely identified by alphanumeric technical identifiers like *R-71-CONDITIONAL-PROPERTIES*

lished in a technical report [7]. In the following, we summarize the classifications of constraint types and constraints for the purpose of our evaluation.

3.1 Classification of Constraint Types according to the Expressivity of Constraint Languages

According to the expressivity of constraint languages, the complete set of constraint types encompasses three not disjoint sets of constraint types:

1. *RDFS/OWL Based*
2. *Constraint Language Based*
3. *SPARQL Based*

RDFS/OWL Based. The modeling languages RDFS and OWL are typically used to formally specify vocabularies. *RDFS/OWL Based* denotes the set of constraint types which can be formulated with RDFS/OWL axioms which we use in terms of constraints with CWA/UNA semantics and without reasoning.[11]

Constraint Language Based. We further distinguish *Constraint Language Based* as the set of constraint types that can be expressed by common classical declarative high-level constraint languages like ShEx, ReSh, and DSP. There is a strong overlap between *RDFS/OWL* and *Constraint Language Based* constraint types as in many cases constraint types are expressible by both classical constraint languages and OWL. SPARQL, however, is considered as a low-level implementation language in this context. In contrast to SPARQL, high-level constraint languages are comparatively easy to understand and constraints can be formulated more concisely. Declarative languages may be placed on top of SPARQL when using it as an implementation language.

SPARQL Based. The set *SPARQL Based* encompasses constraint types that are not expressible by RDFS/OWL or common high-level constraint languages but by plain SPARQL.

3.2 Classification of RDF Constraints according to the Severity of Constraint Violations

A concrete constraint is instantiated from one of the 81 constraint types and is defined for a specific vocabulary. It does not make sense to determine the severity of constraint violations of an entire constraint type, as the severity depends on the individual context and vocabulary. SBE experts determined the default *severity level*[12] for each constraint to indicate how serious the violation of the constraint is. We use the classification system of log messages in software

[11] The entailment regime is to be decided by the implementers. It is our point that reasoning affects validation and that a proper definition of the reasoning to be applied is needed.

[12] The possibility to define severity levels in vocabularies is in itself a requirement (*R-158*).

development like *Apache Log4j 2* [1], the *Java Logging API*,[13] and the *Apache Commons Logging API*[14] as many data practitioners also have experience in software development and software developers intuitively understand these levels. We simplify this commonly accepted classification system and distinguish the three severity levels (1) *informational*, (2) *warning*, and (3) *error*. Violations of *informational* constraints point to desirable but not necessary data improvements to achieve RDF representations which are ideal in terms of syntax and semantics of used vocabularies. *Warnings* are syntactic or semantic problems which typically should not lead to an abortion of data processing. *Errors*, in contrast, are syntactic or semantic errors which should cause the abortion of data processing. Although we provide default severity levels for each constraint, validation environments should enable users to adapt the severity levels of constraints according to their individual needs.

4 Evaluation

In this section, we describe our results based on an automatic constraint checking of a large data set. Despite the large volume of the data set in general, we have to keep in mind that this study only uses data for three vocabularies. As described in Section 2, for other vocabularies there is often not (yet) enough data openly available to draw general conclusions. The three vocabularies, however, are representative, cover different aspects of SBE data, and are also a mixture of widely adopted and accepted well-established vocabularies (QB, SKOS) and a vocabulary under development (DDI-RDF[15]).

4.1 Experimental Setup

On the three vocabularies (DDI-RDF, QB, SKOS), we identified and classified 115 constraints[16] which we implemented for data validation. We ensured that the implementation of the constraints is equally distributed over the classes and vocabularies we have. We then evaluated the data quality of 15,694 data sets (4.26 billion triples) of SBE research data using these 115 constraints, obtained from 33 SPARQL endpoints.

Table 1 lists the number of validated data sets and the overall sizes in terms of triples for each of the vocabularies. We validated, i.a., (1) QB data sets published by the *Australian Bureau of Statistics*, the *European Central Bank*, and the *Organisation for Economic Co-operation and Development*, (2) SKOS thesauri like the *AGROVOC Multilingual agricultural thesaurus*, the *STW Thesaurus for Economics*, and the *Thesaurus for the Social Sciences*, and (3) DDI-RDF data

[13] http://docs.oracle.com/javase/7/docs/api/java/util/logging/Level.html
[14] http://commons.apache.org/proper/commons-logging/
[15] Expected publication at the end of the year 2015
[16] All 115 implemented constraints are online available at: https://github.com/boschthomas/rdf-validation/tree/master/constraints

sets provided by the *Microdata Information System*, the *Data Without Boundaries Discovery Portal*, the *Danish Data Archive*, and the *Swedish National Data Service*. We published the evaluation results for each QB data set in form of one document per SPARQL endpoint.[17]

Table 1: Validated Data Sets for each Vocabulary

Vocabulary	Data Sets	Triples
QB	9,990	3,775,983,610
SKOS	4,178	477,737,281
DDI-RDF	1,526	9,673,055

Since the validation of each of the 81 constraint types can be implemented using SPARQL, we use *SPIN*, a SPARQL-based way to formulate and check constraints, as basis to develop a validation environment to validate RDF data according to constraints expressed by arbitrary constraint languages[18] [4]. The *RDF Validator*[19] can directly be used to validate arbitrary RDF data for the three vocabularies. Additionally, own constraints on any vocabulary can be defined using several constraint languages. The SPIN engine checks for each resource if it satisfies all constraints, which are associated with its assigned classes, and generates a result RDF graph containing information about all constraint violations. There is one SPIN construct template for each constraint type. A SPIN construct template contains a SPARQL CONSTRUCT query which generates constraint violation triples indicating the subject and the properties causing constraint violations and the reason why constraint violations have been raised. A SPIN construct template creates constraint violation triples if all triple patterns within the SPARQL WHERE clause match.

4.2 Evaluation Results

Tables 2 and 3 show the results of the evaluation, more specifically the constraints and the constraint violations, which are caused by these constraints, in percent; whereas the numbers in the first line indicate the absolute amount of constraints and violations. The constraints and their raised violations are grouped by vocabulary, which type of language the constraint types are formulated with, and their severity level. The numbers of validated triples and data sets differ between the vocabularies as we validated 3.8 billion QB, 480 million SKOS, and 10 million DDI-RDF triples. To be able to formulate findings which

[17] Online available at: https://github.com/boschthomas/rdf-validation/tree/master/evaluation/data-sets/data-cube
[18] Constraint language implementations online available at: https://github.com/boschthomas/rdf-validation/tree/master/SPIN
[19] Online demo available at: http://purl.org/net/rdfval-demo, source code online available at: https://github.com/boschthomas/rdf-validator

apply for all vocabularies, we only use normalized relative values representing the percentage of constraints and violations belonging to the respective sets.

There is a strong overlap between *RDFS/OWL* and *Constraint Language Based* constraint types as in many cases constraint types are expressible by RDFS/OWL and classical constraint languages. This is the reason why the percentage values of constraints and violations grouped by the classification of constraint types according to the expressivity of constraint languages do not accumulate to 100%.

Table 2: Constraints and Constraint Violations (1)

	DDI-RDF		**QB**	
	C	CV	C	CV
	78	3,575,002	20	45,635,861
SPARQL	29.5	34.7	**60.0**	**100.0**
CL	**64.1**	**65.3**	40.0	0.0
RDFS/OWL	**66.7**	**65.3**	40.0	0.0
info	**56.4**	**52.6**	0.0	0.0
warning	11.5	29.4	15.0	**99.8**
error	32.1	18.0	**85.0**	0.3

C (constraints), CV (constraint violations)

Table 3: Constraints and Constraint Violations (2)

	SKOS		***Total***	
	C	CV	C	CV
	17	5,540,988	115	54,751,851
SPARQL	**100.0**	**100.0**	**63.2**	**78.2**
CL	0.0	0.0	34.7	21.8
RDFS/OWL	0.0	0.0	35.6	21.8
info	**70.6**	41.2	**42.3**	31.3
warning	29.4	**58.8**	18.7	**62.7**
error	0.0	0.0	**39.0**	6.1

C (constraints), CV (constraint violations)

4.3 Legend

In this sub-section, we describe how the tables in this paper should be read. Table 4 gives an overview over the symbols used in subsequent tables of the detailed evaluation.

Symbol	Description
✓	Validation Successful (without any constraint violation)
X	Constraint Violations
>X	Poor Performance/Scaling
✗	Very Poor Performance/Scaling
(!)	Not Yet Implemented Constraint
(X)	The validation of X data sets could not be finished, due to SPARQL endpoints' technical restrictions (e.g., defined timeouts).
*	default severity level *informational*
**	default severity level *warning*
***	default severity level *error*

Table 4: Legend

- **Constraint Violations.** When constraints are violated, X indicates the number of raised constraint violation triples.
- **Poor Performance/Scaling.** The performance of the implementation of the underlying SPARQL CONSTRUCT query is too poor to get all resulting constraint violation triples. Therefore, a limit of X result constraint violation triples is set. It is likely that there are more than X constraint violations. Although the result set contains not the whole set of raised constraint violation triples, the constraint can be used as an indicator if there is data not conforming to the constraint and to resolve constraint violations step by step. As part of future work, the performance will be improved.
- **Very Poor Performance/Scaling.** The performance of the implementation of the underlying SPARQL CONSTRUCT query is too poor to get any results, even though a limit of result constraint violation triples is set. As part of future work, the performance will be improved.

5 Evaluation of Metadata on Unit-Record Data Sets (DDI-RDF)

In this section, the quality of the metadata on unit-record data sets (DDI-RDF) is evaluated by validating appropriate RDF constraints assigned to several RDF constraint types. First, we give an overview on the evaluated data sets and finally we provide details about the evaluation.

5.1 Data Sets Overview

Tables 5 and 7 give an overview on the evaluated DDI-RDF data sets, their abbreviations, and publicly available SPARQL endpoints. Table 6 comprehends

the number of triples, data sets, and instances of multiple vocabulary-specific classes.

Abbr.	**DDI-RDF Data Sets**
Missy	*Microdata Information System*[20]
DwB	*DwB Discovery Portal*[21]
DDA-SND	*DDI-RDF*[22]
	provided by the *Danish Data Archive (DDA)*[23] and Swedish National Data Service (SND)[24]

Table 5: DDI-RDF Data Sets Abbreviations

	Counts									
Data Sets	**triples**	**disco:StudyGroup**	**disco:Study**	**disco:LogicalDataSet**	**disco:Universe**	**disco:Variable**	**disco:Question**	**disco:SummaryStatistics**	**disco:CategoryStatistics**	**skos:Concept**
Missy	5,068,838	6	45	159	1,125	21,040	0	0	0	147,193
DwB	2,332,802	0	1,387	1,367	2,796	446,806	0	0	0	0
DDA-SND	2,271,415	0	1,490	0	10,188	80,070	139,237	0	0	290,963
Total	9,673,055			1,526						

Table 6: DDI-RDF Data Sets Overview

[20] http://www.gesis.org/missy/eu/missy-home
[21] http://dwb-dev.nsd.uib.no/portal
[22] http://ddi-rdf.borsna.se/
[23] http://samfund.dda.dk/dda/default-en.asp
[24] http://snd.gu.se/en

Data Sets	SPARQL Endpoint
Missy	http://svko-missy:8181/openrdf-workbench/repositories/native-java-store/summary
DwB	http://dwb-dev.nsd.uib.no/sparql
DDA-SND	http://ddi-rdf.borsna.se/endpoint/

Table 7: DDI-RDF SPARQL Endpoints

5.2 Detailed Evaluation

In this sub-section, we give details about the evaluation in form of diverse tables containing the number of constraint violations per evaluated data set and constraint of particular constraint types.

	Data Sets		
Existential Quantifications (1)	*Missy*	*DwB*	*DDA-SND*
*DISCO-C-EXISTENTIAL-QUANTIFICATIONS-01****	✓	✓	✓
*DISCO-C-EXISTENTIAL-QUANTIFICATIONS-02****	7	17	1,490
*DISCO-C-EXISTENTIAL-QUANTIFICATIONS-03**	✓	✓	✓
*DISCO-C-EXISTENTIAL-QUANTIFICATIONS-04**	11,021	445,381	62,260
*DISCO-C-EXISTENTIAL-QUANTIFICATIONS-05**	✓	✓	139,237
*DISCO-C-EXISTENTIAL-QUANTIFICATIONS-06**	12	1,367	✓
*DISCO-C-EXISTENTIAL-QUANTIFICATIONS-07**	6	✓	✓
*DISCO-C-EXISTENTIAL-QUANTIFICATIONS-08**	45	1,387	1,490
*DISCO-C-EXISTENTIAL-QUANTIFICATIONS-09**	6	✓	✓
*DISCO-C-EXISTENTIAL-QUANTIFICATIONS-10**	45	1,387	1,490

Table 8: Evaluation of DDI-RDF Data Sets - Existential Quantifications (1)

	Data Sets		
Existential Quantifications (2)	*Missy*	*DwB*	*DDA-SND*
*DISCO-C-EXISTENTIAL-QUANTIFICATIONS-11**	6	✓	✓
*DISCO-C-EXISTENTIAL-QUANTIFICATIONS-12**	6	✓	✓
*DISCO-C-EXISTENTIAL-QUANTIFICATIONS-13**	✓	✓	✓
*DISCO-C-EXISTENTIAL-QUANTIFICATIONS-14**	45	1,387	1,490
*DISCO-C-EXISTENTIAL-QUANTIFICATIONS-15**	45	1,387	1,490
*DISCO-C-EXISTENTIAL-QUANTIFICATIONS-16**	✓	✓	✓
*DISCO-C-EXISTENTIAL-QUANTIFICATIONS-17**	159	1,367	✓
*DISCO-C-EXISTENTIAL-QUANTIFICATIONS-18**	159	1,367	✓
*DISCO-C-EXISTENTIAL-QUANTIFICATIONS-19**	✓	✓	✓
*DISCO-C-EXISTENTIAL-QUANTIFICATIONS-20**	✓	1,367	✓

Table 9: Evaluation of DDI-RDF Data Sets - Existential Quantifications (2)

	Data Sets		
Existential Quantifications (3)	*Missy*	*DwB*	*DDA-SND*
DISCO-C-EXISTENTIAL-QUANTIFICATIONS-21[*]	✓	1,367	✓
DISCO-C-EXISTENTIAL-QUANTIFICATIONS-22[*]	✓	✓	✓
DISCO-C-EXISTENTIAL-QUANTIFICATIONS-23[*]	6	✓	✓
DISCO-C-EXISTENTIAL-QUANTIFICATIONS-24[*]	45	1,387	1,490
DISCO-C-EXISTENTIAL-QUANTIFICATIONS-25[*]	45	1,387	1,490
DISCO-C-EXISTENTIAL-QUANTIFICATIONS-26[*]	45	1,387	1,490
DISCO-C-EXISTENTIAL-QUANTIFICATIONS-27[***]	✓	130	1,490
DISCO-C-EXISTENTIAL-QUANTIFICATIONS-28[**]	159	✓	✓
DISCO-C-EXISTENTIAL-QUANTIFICATIONS-29[**]	✓	✓	✓
DISCO-C-EXISTENTIAL-QUANTIFICATIONS-30[**]	✓	✓	✓

Table 10: Evaluation of DDI-RDF Data Sets - Existential Quantifications (3)

Existential Quantifications (4)	Data Sets		
	Missy	*DwB*	*DDA-SND*
*DISCO-C-EXISTENTIAL-QUANTIFICATIONS-31***	159	1,367	✓
*DISCO-C-EXISTENTIAL-QUANTIFICATIONS-32****	✓	✓	✓
*DISCO-C-EXISTENTIAL-QUANTIFICATIONS-33****	✓	✓	✓
*DISCO-C-EXISTENTIAL-QUANTIFICATIONS-34****	✓	✓	✓
*DISCO-C-EXISTENTIAL-QUANTIFICATIONS-35****	✓	✓	✓
*DISCO-C-EXISTENTIAL-QUANTIFICATIONS-36****	✓	✓	✓
*DISCO-C-EXISTENTIAL-QUANTIFICATIONS-37**	18,625	✓	✓
*DISCO-C-EXISTENTIAL-QUANTIFICATIONS-38**	✓	✓	750
*DISCO-C-EXISTENTIAL-QUANTIFICATIONS-39****	✓	✓	✓
*DISCO-C-EXISTENTIAL-QUANTIFICATIONS-40**	✓	✓	139,237

Table 11: Evaluation of DDI-RDF Data Sets - Existential Quantifications (4)

Existential Quantifications (5)	Data Sets		
	Missy	*DwB*	*DDA-SND*
*DISCO-C-EXISTENTIAL-QUANTIFICATIONS-41**	✓	✓	✓
*DISCO-C-EXISTENTIAL-QUANTIFICATIONS-42**	✓	✓	✓
*DISCO-C-EXISTENTIAL-QUANTIFICATIONS-43**	15,733	446,806	80,070
*DISCO-C-EXISTENTIAL-QUANTIFICATIONS-44**	159	✓	✓
*DISCO-C-EXISTENTIAL-QUANTIFICATIONS-45**	6,784	446,806	19,221
*DISCO-C-EXISTENTIAL-QUANTIFICATIONS-46***	11,550	446,806	10,451

Table 12: Evaluation of DDI-RDF Data Sets - Existential Quantifications (5)

	Data Sets		
Conditional Properties	*Missy*	*DwB*	*DDA-SND*
DISCO-C-CONDITIONAL-PROPERTIES-01***	✓	✓	80,070
DISCO-C-CONDITIONAL-PROPERTIES-02**	12	✓	✓
DISCO-C-CONDITIONAL-PROPERTIES-03**	90	✓	2,980
DISCO-C-CONDITIONAL-PROPERTIES-04***	6	✓	✓
DISCO-C-CONDITIONAL-PROPERTIES-05***	45	1,387	1,490
DISCO-C-CONDITIONAL-PROPERTIES-06***	✓	✓	✓

Table 13: Evaluation of DDI-RDF Data Sets - Conditional Properties

	Data Sets		
Provenance	*Missy*	*DwB*	*DDA-SND*
DISCO-C-PROVENANCE-01*	6	✓	✓
DISCO-C-PROVENANCE-02*	45	1,387	1,490
DISCO-C-PROVENANCE-03*	159	1,367	✓
DISCO-C-PROVENANCE-04*	✓	1,367	✓

Table 14: Evaluation of DDI-RDF Data Sets - Provenance

	Data Sets		
Labeling and Documentation	*Missy*	*DwB*	*DDA-SND*
DISCO-C-LABELING-AND-DOCUMENTATION-01[*]	6	✓	✓
DISCO-C-LABELING-AND-DOCUMENTATION-02[*]	45	1,387	1,490
DISCO-C-LABELING-AND-DOCUMENTATION-03[*]	159	1,367	✓
DISCO-C-LABELING-AND-DOCUMENTATION-04[*]	✓	1,367	✓
DISCO-C-LABELING-AND-DOCUMENTATION-05[*]	✓	✓	✓
DISCO-C-LABELING-AND-DOCUMENTATION-06[*]	21,040	446,806	80,070

Table 15: Evaluation of DDI-RDF Data Sets - Labeling and Documentation

	Data Sets		
Data Model Consistency	*Missy*	*DwB*	*DDA-SND*
DISCO-C-DATA-MODEL-CONSISTENCY-01 (!)[***]			
DISCO-C-DATA-MODEL-CONSISTENCY-02 (!)[***]			
DISCO-C-DATA-MODEL-CONSISTENCY-03 (!)[***]			
DISCO-C-DATA-MODEL-CONSISTENCY-04 (!)[***]			
DISCO-C-DATA-MODEL-CONSISTENCY-05[***]	✓	✓	✓
DISCO-C-DATA-MODEL-CONSISTENCY-06 (!)[***]			
DISCO-C-DATA-MODEL-CONSISTENCY-07 (!)[***]			

Table 16: Evaluation of DDI-RDF Data Sets - Data Model Consistency

	Data Sets		
Comparison	*Missy*	*DwB*	*DDA-SND*
*DISCO-C-COMPARISON-VARIABLES-01 (!)***			
*DISCO-C-COMPARISON-VARIABLES-02****	21,040	446,806	80,070
*DISCO-C-COMPARISON-VARIABLES-03 (!)****			
*DISCO-C-COMPARISON-VARIABLES-04**	18,625	✓	✓
*DISCO-C-COMPARISON-VARIABLES-05****	159	✓	✓

Table 17: Evaluation of DDI-RDF Data Sets - Comparison

	Data Sets		
Mathematical Operations	*Missy* *DwB*		*DDA-SND*
*DISCO-C-MATHEMATICAL-OPERATIONS-01 (!)****			
*DISCO-C-MATHEMATICAL-OPERATIONS-02 (!)****			
*DISCO-C-MATHEMATICAL-OPERATIONS-03 (!)****			
*DISCO-C-MATHEMATICAL-OPERATIONS-04 (!)****			
*DISCO-C-MATHEMATICAL-OPERATIONS-05 (!)****			

Table 18: Evaluation of DDI-RDF Data Sets - Mathematical Operations

	Data Sets	
Language Tags	*Missy DwB*	*DDA-SND*
*DISCO-C-LANGUAGE-TAG-MATCHING-01 (!)**		
*DISCO-C-LANGUAGE-TAG-CARDINALITY-01 (!)**		
*DISCO-C-LANGUAGE-TAG-CARDINALITY-02 (!)**		
*DISCO-C-LANGUAGE-TAG-CARDINALITY-03 (!)**		

Table 19: Evaluation of DDI-RDF Data Sets - Language Tags

	Data Sets	
Aggregation	*Missy DwB*	*DDA-SND*
*DISCO-C-AGGREGATION-01 (!)**		
*DISCO-C-AGGREGATION-02 (!)**		
*DISCO-C-AGGREGATION-03 (!)**		
*DISCO-C-AGGREGATION-04 (!)**		
*DISCO-C-AGGREGATION-05 (!)**		
*DISCO-C-AGGREGATION-06 (!)**		
*DISCO-C-AGGREGATION-07 (!)**		

Table 20: Evaluation of DDI-RDF Data Sets - Aggregation

DDI-RDF Constraints	Data Sets		
	Missy	*DwB*	*DDA-SND*
DISCO-C-ALLOWED-VALUES-01***	✓	✓	✓
DISCO-C-LITERAL-RANGES-01***	✓	✓	✓
DISCO-C-INVERSE-FUNCTIONAL-PROPERTIES-01***	✓	✓	✓
DISCO-C-INVERSE-FUNCTIONAL-PROPERTIES-02***	✓	✓	✓
DISCO-C-CLASS-SPECIFIC-PROPERTY-RANGE-01***	✓	✓	✓
DISCO-C-MEMBERSHIP-IN-CONTROLLED-VOCABULARIES-01***	✓	✓	✗
DISCO-C-LITERAL-VALUE-COMPARISON-01***	✓	1,299	✓
DISCO-C-CONTEXT-SPECIFIC-VALID-PROPERTIES-01*	21,038	✓	✓
DISCO-C-DATA-PROPERTY-FACETS-01**	✓	✓	✓
DISCO-C-DATA-PROPERTY-FACETS-02**	✓	✓	✓

Table 21: Evaluation of DDI-RDF Data Sets - DDI-RDF Constraints (1)

	Data Sets		
DDI-RDF Constraints	*Missy*	*DwB*	*DDA-SND*
DISCO-C-VALUE-IS-VALID-FOR-DATATYPE-01***	30	6,932	✓
DISCO-C-VALUE-IS-VALID-FOR-DATATYPE-02***	✓	✓	✓
DISCO-C-SUBSUMPTION-01 (!)$^{***\mathrm{B}}$			
DISCO-C-CLASS-EQUIVALENCE-01 (!)*			
DISCO-C-SUB-PROPERTIES-01 (!)***			
DISCO-C-PROPERTY-DOMAIN-01 (!)***			
DISCO-C-PROPERTY-RANGES-01 (!)***			
DISCO-C-INVERSE-OBJECT-PROPERTIES-01 (!)***			
DISCO-C-INVERSE-OBJECT-PROPERTIES-02 (!)***			
DISCO-C-INVERSE-OBJECT-PROPERTIES-03 (!)***			
DISCO-C-DISJOINT-PROPERTIES-01 (!)***			

Table 22: Evaluation of DDI-RDF Data Sets - DDI-RDF Constraints (2)

DDI-RDF Constraints	Data Sets: *Missy DwB*	Data Sets: *DDA-SND*
DISCO-C-ASYMMETRIC-OBJECT-PROPERTIES-01 (!)***		
DISCO-C-IRREFLEXIVE-OBJECT-PROPERTIES-01 (!)***		
DISCO-C-CLASS-SPECIFIC-IRREFLEXIVE-OBJECT-PROPERTIES-01 (!)***		
DISCO-C-CLASS-SPECIFIC-IRREFLEXIVE-OBJECT-PROPERTIES-02 (!)***		
DISCO-C-DISJOINT-CLASSES-01 (!)***		
DISCO-C-EQUIVALENT-PROPERTIES-01 (!)*		
DISCO-C-LITERAL-PATTERN-MATCHING-01 (!)*		
DISCO-C-DISJUNCTION-01 (!)***		
DISCO-C-UNIVERSAL-QUANTIFICATIONS-01 (!)***		
DISCO-C-MINIMUM-QUALIFIED-CARDINALITY-RESTRICTIONS-01 (!)***		

Table 23: Evaluation of DDI-RDF Data Sets - DDI-RDF Constraints (3)

DDI-RDF Constraints	Data Sets: *Missy DwB*	Data Sets: *DDA-SND*
DISCO-C-MAXIMUM-QUALIFIED-CARDINALITY-RESTRICTIONS-01 (!)***		
DISCO-C-EXACT-QUALIFIED-CARDINALITY-RESTRICTIONS-01 (!)***		
DISCO-C-CONTEXT-SPECIFIC-EXCLUSIVE-OR-OF-PROPERTY-GROUPS-01 (!)*		
DISCO-C-IRI-PATTERN-MATCHING-01 (!)*		
DISCO-C-ORDERING-01 (!)*		
DISCO-C-ORDERING-02 (!)*		
DISCO-C-ORDERING-03 (!)*		
DISCO-C-STRING-OPERATIONS-01 (!)*		
DISCO-C-CONTEXT-SPECIFIC-VALID-CLASSES-01 (!)*		
DISCO-C-CONTEXT-SPECIFIC-VALID-PROPERTIES-01 (!)*		

Table 24: Evaluation of DDI-RDF Data Sets - DDI-RDF Constraints (4)

DDI-RDF Constraints	Data Sets: *Missy DwB*	Data Sets: *DDA-SND*
DISCO-C-DEFAULT-VALUES-01 (!)[*]		
DISCO-C-WHITESPACE-HANDLING-01 (!)[*]		
DISCO-C-HTML-HANDLING-01 (!)[*]		
DISCO-C-HTML-HANDLING-02 (!)[*]		
DISCO-C-RECOMMENDED-PROPERTIES-01 (!)[*]		
DISCO-C-HANDLE-RDF-COLLECTIONS-01 (!)[*]		
DISCO-C-HANDLE-RDF-COLLECTIONS-02 (!)[*]		
DISCO-C-USE-SUB-SUPER-RELATIONS-IN-VALIDATION-01 (!)[*]		
DISCO-C-USE-SUB-SUPER-RELATIONS-IN-VALIDATION-02 (!)[*]		
DISCO-C-STRUCTURE-01 (!)[***]		

Table 25: Evaluation of DDI-RDF Data Sets - DDI-RDF Constraints (5)

DDI-RDF Constraints	Data Sets: *Missy DwB*	Data Sets: *DDA-SND*
DISCO-C-VOCABULARY-01 (!)[***]		
DISCO-C-HTTP-URI-SCHEME-VIOLATION (!)[***]		

Table 26: Evaluation of DDI-RDF Data Sets - DDI-RDF Constraints (6)

6 Evaluation of Metadata and Data of Aggregated Data Sets (QB)

In this section, the quality of the metadata on aggregated data (QB) data sets and of the data sets themselves is evaluated by validating appropriate RDF con-

straints assigned to several RDF constraint types. First, we we give an overview on the evaluated data sets and finally we provide details about the evaluation.

6.1 Data Sets Overview

There are websites giving an overview on available QB data sets[25]. Tables 27 and 29 give an overview on the evaluated QB data sets, their abbreviations, and publicly available SPARQL endpoints. Table 28 comprehends the number of triples, data sets, and instances of multiple vocabulary-specific classes.

Abbr.	**QB Data Sets**
ECB	*European Central Bank*[26]
UIS	*UNESCO Institute for Statistics*[27]
IMF	*International Monetary Fund*[28]
BFS	*Bundesamt für Statistik - Swiss Federal Statistics*[29]
FAO	*Food and Agriculture Organization of the United Nations*[30]
WB	*World Bank*[31]
FRB	*Federal Reserve Board*[32]
TI	*Transparency International*[33]
OECD	*Organisation for Economic Co-operation and Development*[34]
BIS	*Bank for International Settlements*[35]
ABS	*Australian Bureau of Statistics*[36]
IEEE-VIS	*IEEE VIS Source Data*
ACORN-SAT	*Australian Climate Observations Reference Network - Surface Air Temperature Dataset*
HDP	*HealthData.gov Platform (HDP) on the Semantic Web*
Eurostat	*The Eurostat Linked Data* (SPARQL endpoint unavailable)
Asturias	*Nomenclator Asturias* (SPARQL endpoint unavailable!)
ISTAT	*ISTAT Immigration (LinkedOpenData.it)* (SPARQL endpoint unavailable)
ICANE	*Statistical Office of Cantabria (Instituto Cántabro de Estadística, ICANE)* (SPARQL endpoint unavailable)
EE-2009	*European Election Results 2009* (SPARQL endpoint unavailable)
EU-B	*Standard Eurobarometer* (SPARQL endpoint unavailable)
ECB-S	*European Central Bank Statistics (PublicData.eu)* (SPARQL endpoint unavailable)
CPV-2008	*Common Procurement Vocabulary (CPV) 2008* (SPARQL endpoint unavailable)
CPV-2003	*Common Procurement Vocabulary (CPV) 2003* (SPARQL endpoint unavailable)

Table 27: QB Data Sets Abbreviations

[25] http://270a.info/; http://datahub.io/de/dataset?tags=format-qb; http://ontologycentral.com/

[26] http://www.ecb.europa.eu/home/html/index.en.html
[27] http://www.uis.unesco.org/Pages/default.aspx
[28] http://www.imf.org/external/index.htm
[29] http://www.bfs.admin.ch/
[30] http://www.fao.org/home/en/
[31] http://www.worldbank.org/
[32] http://www.federalreserve.gov/
[33] http://www.transparency.org/
[34] http://www.oecd.org/
[35] http://www.bis.org/
[36] http://abs.gov.au/

	Counts				
Data Sets	**triples**	**qb:DataSet**	**qb:DataStructureDefinition**	**qb:Observation**	**qb:Slice**
ECB	468,899,474	55	46	>11,000,000	428,698
UIS	10,400,534	5	5	1,437,651	0
IMF	35,688,446	4	8	3,603,719	0
BFS	1,533,743	0	0	8	0
FAO	53,000,000	10	10	>7,100,000	0
WB	174,006,552	9,466	59	>17,000,000	0
FRB	185,266,900	49	98	>9,500,000	0
TI	52,233	6	6	3,928	0
OECD	304,995,160	136	140	>12,000,000	0
BIS	54,197,482	6	12	3,606,466	47,914
ABS	2,357,400,000	253	257	>11,000,000	0
IEEE-VIS	19,935,340	0	0	1,350	0
ACORN-SAT	98,381,319	0	4	0	0
HDP	12,226,427	0	0	0	0
Total	3,775,983,610	9,990			

Table 28: QB Data Sets Overview

Data Sets	SPARQL Endpoints
ECB	http://ecb.270a.info/sparql
UIS	http://uis.270a.info/sparql
IMF	http://imf.270a.info/sparql
BFS	http://bfs.270a.info/sparql
FAO	http://fao.270a.info/sparql
WB	http://worldbank.270a.info/sparql
FRB	http://frb.270a.info/sparql
TI	http://transparency.270a.info/sparql
OECD	http://oecd.270a.info/sparql
BIS	http://bis.270a.info/sparql
ABS	http://abs.270a.info/sparql
ACORN-SAT	http://lab.environment.data.gov.au/sparql
HDP	http://healthdata.tw.rpi.edu/sparql

Table 29: QB SPARQL Endpoints

6.2 Detailed Evaluation

In this sub-section, we give details about the evaluation in form of diverse tables containing the number of constraint violations per evaluated data set and constraint of particular constraint types.

	Data Sets						
Data Model Consistency	*ECB*	*UIS*	*IMF*	*BFS*	*FAO*	*WB*	*FRB*
*DATA-MODEL-CONSISTENCY-01***	✓ (2)	✓	✓	✓	✓	✓	✓
*DATA-MODEL-CONSISTENCY-02****	✓ (2)	✓	✓	✓	✓	✓	✓
*DATA-MODEL-CONSISTENCY-03****	✓ (2)	✓	✓	✓	✓	✓	✓
*DATA-MODEL-CONSISTENCY-04****	✓ (6)	✓	✓	✓	✓	✓	14,372
*DATA-MODEL-CONSISTENCY-05***	1,198,352 (50)	✗	✗	✓	✗	✓	16,175,814 (42)
*DATA-MODEL-CONSISTENCY-06****	✓ (2)	✓	✓	✓	✓	✓	✓
*DATA-MODEL-CONSISTENCY-07****	✓ (9)	✓	99,091	✓	✓	✓	✓ (1)
*DATA-MODEL-CONSISTENCY-08****	✓ (2)	✓	✓	✓	✓	✓	✓
*DATA-MODEL-CONSISTENCY-09****	✓ (2)	✓	✓	✓	✓	✓	✓
*DATA-MODEL-CONSISTENCY-10**** (!)	-	-	-	-	-	-	-
*DATA-MODEL-CONSISTENCY-11***	6,511 (10)	✓	✓	✓	✓	✓	✓

Table 30: Evaluation of QB Data Sets - Data Model Consistency (1)

	Data Sets						
Data Model Consistency	*TI*	*OECD*	*BIS*	*ABS*	*IEEE-VIS*	*ACORN-SAT*	*HDP*
*DATA-MODEL-CONSISTENCY-01***	✓	✓	✓	✓	✓	✓	✓
*DATA-MODEL-CONSISTENCY-02****	✓	✓	✓	✓	✓	8	✓
*DATA-MODEL-CONSISTENCY-03****	✓	✓	✓	✓	✓	✓	✓
*DATA-MODEL-CONSISTENCY-04****	✓	✓	✓	✓ (6)	✓	✓	✓
*DATA-MODEL-CONSISTENCY-05***	✓	21,142,838 (116)	✗	6,997,098 (246)	✓	✓	✓
*DATA-MODEL-CONSISTENCY-06****	✓	✓	✓	✓	✓	✓	✓
*DATA-MODEL-CONSISTENCY-07****	✓	✓	✓	✓ (8)	✓	✓	✓
*DATA-MODEL-CONSISTENCY-08****	✓	✓	✓	✓	✓	✓	✓
*DATA-MODEL-CONSISTENCY-09****	✓	✓	✓	✓	✓	✓	✓
*DATA-MODEL-CONSISTENCY-10**** (!)	-	-	-	-	-	-	-
*DATA-MODEL-CONSISTENCY-11***	✓	✓	✓	✓	✓	✓	✓

Table 31: Evaluation of QB Data Sets - Data Model Consistency (2)

	Data Sets													
Existential Quantifications	*ECB*	*UIS*	*IMF*	*BFS*	*FAO*	*WB*	*FRB*	*TI*	*OECD*	*BIS*	*ABS*	*IEEE-VIS*	*ACORN-SAT*	*HDP*
*EXISTENTIAL-QUANTIFICATIONS-01****	9	✓	11	7	8	77	8	9	7	8	7	✓	✓	✓
*EXISTENTIAL-QUANTIFICATIONS-02****	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
*EXISTENTIAL-QUANTIFICATIONS-03****	✓	✓	✓	✓	✓	59	✓	6	✓	✓	✓	✓	4	✓
*EXISTENTIAL-QUANTIFICATIONS-04****	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓

Table 32: Evaluation of QB Data Sets - Existential Quantifications

Cardinality Restrictions	Data Sets									
	ECB	*UIS*	*IMF*	*BFS*	*FAO*	*WB*	*FRB*	*TI*	*OECD*	*BIS*
MINIMUM-QUALIFIED-CARDINALITY-RESTRICTIONS-01 (!)***	-	-	-	-	-	-	-	-	-	-
MINIMUM-QUALIFIED-CARDINALITY-RESTRICTIONS-02***	✗	118	8	8	30	✓	30	✓	✗	12
MAXIMUM-QUALIFIED-CARDINALITY-RESTRICTIONS-01***	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
EXACT-UNQUALIFIED-CARDINALITY-RESTRICTIONS-01***	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
EXACT-QUALIFIED-CARDINALITY-RESTRICTIONS-02***	✓	✓	✓	✓	✓	1	✓	✓	✓	✓

Table 33: Evaluation of QB Data Sets - Cardinality Restrictions (1)

Cardinality Restrictions	Data Sets			
	ABS	*IEEE-VIS*	*ACORN-SAT*	*HDP*
MINIMUM-QUALIFIED-CARDINALITY-RESTRICTIONS-01 (!)***	-	-	-	-
MINIMUM-QUALIFIED-CARDINALITY-RESTRICTIONS-02***	✗	1,350	✓	✓
MAXIMUM-QUALIFIED-CARDINALITY-RESTRICTIONS-01***	✓ (2)	✓	✓	✓
EXACT-UNQUALIFIED-CARDINALITY-RESTRICTIONS-01***	✓	✓	✓	✓
EXACT-QUALIFIED-CARDINALITY-RESTRICTIONS-02***	✓	✓	✓	✓

Table 34: Evaluation of QB Data Sets - Cardinality Restrictions (2)

	Data Sets													
Structure	*ECB*	*UIS*	*IMF*	*BFS*	*FAO*	*WB*	*FRB*	*TI*	*OECD*	*BIS*	*ABS*	*IEEE-VIS*	*ACORN-SAT*	*HDP*
STRUCTURE-01***	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
STRUCTURE-02***	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓

Table 35: Evaluation of QB Data Sets - Structure

	Data Sets													
Constraints	*ECB*	*UIS*	*IMF*	*BFS*	*FAO*	*WB*	*FRB*	*TI*	*OECD*	*BIS*	*ABS*	*IEEE-VIS*	*ACORN-SAT*	*HDP*
PROPERTY-DOMAIN-01 (!)***														
PROPERTY-RANGES-01 (!)***														
DISJOINT-PROPERTIES-01 (!)***														
DISJOINT-CLASSES-01 (!)***														
EQUIVALENT-PROPERTIES-01 (!)*														
UNIVERSAL-QUANTIFICATIONS-01 (!)***														
MEMBERSHIP-IN-CONTROLLED-VOCABULARIES-01 (!)***														
CONTEXT-SPECIFIC-VALID-CLASSES-01 (!)*														
CONTEXT-SPECIFIC-VALID-PROPERTIES-01 (!)*														
RECOMMENDED-PROPERTIES-01 (!)*														
VALUE-IS-VALID-FOR-DATATYPE-01 (!)***														
VOCABULARY-01 (!)***														

Table 36: Evaluation of QB Data Sets - Constraints (1)

Constraints	Data Sets											
	ECB	*UIS*	*IMF*	*BFS*	*FAO*	*WB*	*FRB*	*TI*	*OECD*	*BIS*	*ABS*	*IEEE-VIS*	*ACORN-SAT*	*HDP*
*HTTP-URI-SCHEME-VIOLATION (!)***														

Table 37: Evaluation of QB Data Sets - Constraints (2)

7 Evaluation of Metadata on Thesauri (SKOS)

In this section, the quality of the metadata on thesauri (SKOS) is evaluated by validating appropriate RDF constraints assigned to several RDF constraint types. First, we give an overview on the evaluated thesauri and finally we provide details about the evaluation.

7.1 Data Sets Overview

There is a website giving an overview on available SKOS data sets[37] and another one giving an overview on available thesauri[38]. Tables 38 and 40 give an overview on the evaluated thesauri, their abbreviations, and publicly available SPARQL endpoints. Table 39 comprehends the number of triples, data sets, and instances of multiple vocabulary-specific classes.

[37] http://datahub.io/de/dataset?tags=format-skos
[38] http://datahub.io/de/dataset?tags=thesaurus

Abbr.	Thesauri
TheSoz	*Thesaurus for the Social Sciences*[39]
STW	*Thesaurus for Economics*[40]
AGROVOC	*AGROVOC Multilingual agricultural thesaurus*[41]
UNESCO	*UNESCO Thesaurus*[42]
TGN	*The Getty Thesaurus of Geographic Names*[43]
EARTh	*Environmental Applications Reference Thesaurus*[44]
ODT	*Open Data Thesaurus*[45]
SLD	*Spanish Linguistic Datasets*[46]
SSWT	*Social Semantic Web Thesaurus*[47]
GBA-GU	*Thesaurus of the Geological Survey of Austria (GBA) - Geology Unit*[48]
GBA-GTS	*Thesaurus of the Geological Survey of Austria (GBA) - Geologic Time Scale*[49]
GBA-L	*Thesaurus of the Geological Survey of Austria (GBA) - Lithology*[50]
GBA-LU	*Thesaurus of the Geological Survey of Austria (GBA) - Lithotectonic Unit*[51]
GEMET	*GEneral Multilingual Environmental Thesaurus*[52]
EuroVoc	*EuroVoc*[53]
CECCT	*Clean Energy and Climate Change Thesaurus*[54]

Table 38: Thesauri Abbreviations

[39] http://www.ecb.europa.eu/home/html/index.en.html
[40] http://zbw.eu/stw/versions/latest/about
[41] http://202.45.139.84:10035/catalogs/fao/repositories/agrovoc
[42] http://skos.um.es/sparql/
[43] http://vocab.getty.edu/sparql
[44] http://linkeddata.ge.imati.cnr.it/resource/EARTh/
[45] http://vocabulary.semantic-web.at/PoolParty/wiki/OpenData
[46] http://linguistic.linkeddata.es
[47] http://vocabulary.semantic-web.at/PoolParty/wiki/semweb
[48] http://resource.geolba.ac.at/
[49] http://resource.geolba.ac.at/
[50] http://resource.geolba.ac.at/
[51] http://resource.geolba.ac.at/
[52] http://www.eionet.europa.eu/gemet/
[53] http://open-data.europa.eu/de/data/dataset/eurovoc
[54] http://data.reegle.info/thesaurus/guide

	Counts						
Thesauri	**triples**	**skos:ConceptScheme**	**sko:Concept**	**skos:broader**	**skos:narrower**	**skos:hasTopConcept**	**skos:inScheme**
TheSoz	439,153	1	8,426	13,705	13,706	0	48,529
STW	221,668	1	13,468	13,732	13732	7	13,180
AGROVOC	6,080,477	1	32,310	33,507	33,507	25	32,310
UNESCO	288,346	9	26,714	20,028	20,028	607	32,009
TGN	16,112,321	8	2,898,775	0	0	0	1,453,767
EARTh	9,287,364	11	295,375	288,208	93,827	479	295,376
ODT	3,290	6	108	93	93	30	0
SLD	7,629,211	0	31,195	0	0	0	0
SSWT	64,698	9	2,127	2,300	2,301	38	0
GBA-GU	25,718	3	878	1,005	1,005	14	0
GBA-GTS	7,875	3	213	208	208	5	0
GBA-L	9,317	1	249	249	249	4	0
GBA-LU	9,504	3	364	359	359	7	0
GEMET	372,889,229	3,680	414,659	62,193	21,685	30,806	409,290
EuroVoc	64,477,774	439	79,557	6,922	0	532	14,428
CECCT	191,336	3	3,419	3,761	3,762	28	0
Total	477,737,281	4,178					

Table 39: Thesauri Overview

Thesauri	SPARQL Endpoints
TheSoz	http://lod.gesis.org/thesoz/sparql
STW	http://zbw.eu/beta/sparql/stw/query
AGROVOC	http://202.45.139.84:10035/catalogs/fao/repositories/agrovoc
UNESCO	http://skos.um.es/sparql/
TGN	http://vocab.getty.edu/
EARTh	http://linkeddata.ge.imati.cnr.it:8890/sparql
ODT	http://vocabulary.semantic-web.at/PoolParty/sparql/OpenData
SLD	http://linguistic.linkeddata.es/sparql
SSWT	http://vocabulary.semantic-web.at/PoolParty/sparql/semweb
GBA-GU	http://resource.geolba.ac.at/PoolParty/sparql/GeologicUnit
GBA-GTS	http://resource.geolba.ac.at/PoolParty/sparql/GeologicTimeScale
GBA-L	http://resource.geolba.ac.at/PoolParty/sparql/lithology
GBA-LU	http://resource.geolba.ac.at/PoolParty/sparql/tectonicunit
GEMET	http://semantic.eea.europa.eu/sparql
EuroVoc	http://open-data.europa.eu/de/linked-data
CECCT	http://poolparty.reegle.info/PoolParty/sparql/glossary

Table 40: Thesauri SPARQL Endpoints

7.2 Detailed Evaluation

In this sub-section, we give details about the evaluation in form of diverse tables containing the number of constraint violations per evaluated data set and constraint of particular constraint types.

	Data Sets									
Data Model Consistency	*TheSoz*	*STW*	*AGROVOC*	*TGN*	*UNESCO*	*ODT*	*SSWT*	*GBA-GU*	*GBA-GTS*	*GBA-L*	*GBA-LU*	*CECCT*
*DATA-MODEL-CONSISTENCY-01 (!)**												
*DATA-MODEL-CONSISTENCY-02 (!)**												
*DATA-MODEL-CONSISTENCY-03 (!)**												

Table 41: Thesauri Evaluation - Data Model Consistency (1)

	Data Sets
Data Model Consistency	*EARTh* *GEMET* *Euro Voc* *SLD*
*DATA-MODEL-CONSISTENCY-01 (!)**	
*DATA-MODEL-CONSISTENCY-02 (!)**	
*DATA-MODEL-CONSISTENCY-03 (!)**	

Table 42: Thesauri Evaluation - Data Model Consistency (2)

	Data Sets											
Labeling and Documentation	*TheSoz*	*STW*	*AGROVOC*	*TGN*	*UNESCO*	*ODT*	*SSWT*	*GBA-GU*	*GBA-GTS*	*GBA-L*	*GBA-LU*	*CECCT*
*LABELING-AND-DOCUMENTATION-01**	8,426	11,508	19,829	1,110	✗	36	1,475	5	2	✓	107	486
*LABELING-AND-DOCUMENTATION-02**	>1	✗	>100	287	✗	✓	✓	✓	✓	✓	✓	✓
*LABELING-AND-DOCUMENTATION-03**	✓	✓	1	14,114	✗	✓	✓	1	✓	✓	1	✓
*LABELING-AND-DOCUMENTATION-04 (!)**												
*LABELING-AND-DOCUMENTATION-05**	✓	✓	4	✓	1	2	2	1	✓	✓	✓	7
*LABELING-AND-DOCUMENTATION-06**	975,340	✓	✓	2	✓	✓	✓	✓	✓	✓	✓	✓

Table 43: Thesauri Evaluation - Labeling and Documentation (1)

	Data Sets			
Labeling and Documentation	*EARTh*	*GEMET*	*EuroVoc*	*SLD*
LABELING-AND-DOCUMENTATION-01[*]	264,687	✗	54,911	31,195
LABELING-AND-DOCUMENTATION-02[*]	✗	✗	✗	✓
LABELING-AND-DOCUMENTATION-03[*]	2	✗	55,556	31,195
LABELING-AND-DOCUMENTATION-04 (!)[*]				
LABELING-AND-DOCUMENTATION-05[*]	39	✗	✗	978
LABELING-AND-DOCUMENTATION-06[*]	302	46,718	✓	✓

Table 44: Thesauri Evaluation - Labeling and Documentation (2)

	Data Sets											
Structure	*TheSoz*	*STW*	*AGROVOC*	*TGN*	*UNESCO*	*ODT*	*SSWT*	*GBA-GU*	*GBA-GTS*	*GBA-L*	*GBA-LU*	*CECCT*
STRUCTURE-01[**]	1	1,074	✓	✓	1	5	1	✓	✓	✓	✓	✓
STRUCTURE-02 (!)[*]												
STRUCTURE-03[**]	✓	✓	✓	✓	84	✓	✓	✓	✓	✓	✓	✓
STRUCTURE-04[*]	2,906	8,046	726	✓	3,840	12	124	84	256	68	22	2,422
STRUCTURE-05[*]	✓	✓	✓	✓	✗	90	5,150	✓	✓	✓	✓	9,864
STRUCTURE-06[*]	1,457	37	✓	✓	✗	✓	4	1	1	64	✓	136
STRUCTURE-07[**]	40	5,370	✓	✓	✗	✓	✓	✓	✓	✓	✓	✓
STRUCTURE-08 (!)[***]												
STRUCTURE-09[*]	7,897	19,844	99	✓	552	2	16	26	✓	✓	✓	82
STRUCTURE-10[**]	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓

Table 45: Thesauri Evaluation - Structure (1)

	Data Sets			
Structure	*EARTh*	*GEMET*	*EuroVoc*	*SLD*
*STRUCTURE-01***	18,240	✗	55,757	31,195
*STRUCTURE-02 (!)**				
*STRUCTURE-03***	39	4,244	✓	✓
*STRUCTURE-04**	11,286	74	✓	✓
*STRUCTURE-05**	✓	✗	✓	✓
*STRUCTURE-06**	239,346	✗	13,876	✓
*STRUCTURE-07***	110,015	✗	366,155	155,975
*STRUCTURE-08 (!)****				
*STRUCTURE-09**	107,195	32	✓	✓
*STRUCTURE-10***	27	2,122	✓	✓

Table 46: Thesauri Evaluation - Structure (2)

	Data Sets										
Language Tag Cardinality	*TheSoz*	*STW*	*AGROVOC*	*TGN*	*UNESCO ODT*	*SSWT*	*GBA-GU*	*GBA-GTS*	*GBA-L*	*GBA-LU*	*CECCT*
*LANGUAGE-TAG-CARDINALITY-01***	9,435	13,468	98,894	✓	541	10,147	5,117	2,061	1,742	2,272	15,550
*LANGUAGE-TAG-CARDINALITY-02**	8,222	36,936	✗	✓	265	3,627	2,212	635	631	1,253	9,607
*LANGUAGE-TAG-CARDINALITY-03**	8,222	✓	135	✓	✓	✓	✓	✓	✓	✓	✓
*LANGUAGE-TAG-CARDINALITY-04**	✓	476	✗	50	✓	✓	✓	✓	✓	✓	✓

Table 47: Thesauri Evaluation - Language Tag Cardinality (1)

	Data Sets			
Language Tag Cardinality	*EARTh*	*GEMET*	*Euro Voc*	*SLD*
*LANGUAGE-TAG-CARDINALITY-01***	✗	2,318,895	✗	30,781
*LANGUAGE-TAG-CARDINALITY-02**	✗	✗	✗	✗
*LANGUAGE-TAG-CARDINALITY-03**	224,206	✗	✗	31,195
*LANGUAGE-TAG-CARDINALITY-04**	✗	✗	✓	✓

Table 48: Thesauri Evaluation - Language Tag Cardinality (2)

	Data Sets									
Constraints	*TheSoz*	*STW*	*AGROVOC*	*TGN*	*UNESCO*	*ODT*	*SSWT*	*GBA-GU*	*GBA-GTS*	*GBA-L*	*GBA-LU*	*CECCT*
*PROPERTY-DOMAIN-01 (!)****												
*PROPERTY-RANGES-01 (!)****												
*DISJOINT-PROPERTIES-01 (!)****												
*DISJOINT-PROPERTIES-02 (!)****												
*DISJOINT-CLASSES-01 (!)****												
*EQUIVALENT-PROPERTIES-01 (!)**												
*UNIVERSAL-QUANTIFICATIONS-01 (!)****												
*CONTEXT-SPECIFIC-VALID-CLASSES-01 (!)**												
*CONTEXT-SPECIFIC-VALID-PROPERTIES-01 (!)**												
*RECOMMENDED-PROPERTIES-01 (!)**												
*VOCABULARY-01 (!)****												
*HTTP-URI-SCHEME-VIOLATION (!)****												

Table 49: Thesauri Evaluation - Constraints (1)

Constraints	Data Sets: EARTh	GEMET	Euro Voc	SLD
PROPERTY-DOMAIN-01 (!)[***]				
PROPERTY-RANGES-01 (!)[***]				
DISJOINT-PROPERTIES-01 (!)[***]				
DISJOINT-PROPERTIES-02 (!)[***]				
DISJOINT-CLASSES-01 (!)[***]				
EQUIVALENT-PROPERTIES-01 (!)[*]				
UNIVERSAL-QUANTIFICATIONS-01 (!)[***]				
CONTEXT-SPECIFIC-VALID-CLASSES-01 (!)[*]				
CONTEXT-SPECIFIC-VALID-PROPERTIES-01 (!)[*]				
RECOMMENDED-PROPERTIES-01 (!)[*]				
VOCABULARY-01 (!)[***]				
HTTP-URI-SCHEME-VIOLATION (!)[***]				

Table 50: Thesauri Evaluation - Constraints (2)

8 Evaluation of Metadata on Statistical Classifications (XKOS)

As part of future work, the quality of metadata on statistical classifications (XKOS) data sets will be evaluated by validating appropriate RDF constraints assigned to several RDF constraint types.

8.1 Data Sets Overview

Abbr.	Statistical Classifications
NAF	*Nomenclature d'activités française*[55]
PCS	*Nomenclature des Professions et Catégories Socioprofessionnelles*[56]
CJ	*Nomenclature des catégories juridiques*[57]
ISIC	
ISCO	*International Standard Classification of Occupations*

Table 51: Statistical Classifications Abbreviations

Nomenclature d'activités française (NAF) is the French refinement of the *NACE* classification expressed in XKOS having explanatory notes. *Nomenclature des Professions et Catégories Socioprofessionnelles (PCS)* and *Nomenclature des catégories juridiques (CJ)* are French classifications expressed in XKOS. The statistical classification *ISIC* has explanatory notes too.

9 Conclusion

We identified and published by today 81 types of constraints that are required by various stakeholders for data applications. In close collaboration with several domain experts for the social, behavioral, and economic sciences (SBE), we formulated and implemented 115 constraints on three different vocabularies (DDI-RDF, QB, and SKOS) and classified them according to their severity level and whether their type is expressible by different types of constraint languages - RDFS/OWL, high-level constraint languages, and SPARQL. Using these constraints, we evaluated the data quality of 15,694 data sets (4.26 billion triples) of research data for the SBE sciences obtained from 33 SPARQL endpoints.

References

1. Apache Software Foundation. Apache Log4j 2 v. 2.3 User's Guide. Technical report, Apache Software Foundation, May 2015. http://logging.apache.org/log4j/2.x/log4j-users-guide.pdf.
2. Franz Baader, Diego Calvanese, Deborah L. McGuinness, Daniele Nardi, and Peter F. Patel-Schneider, editors. *The Description Logic Handbook: Theory, Implementation, and Applications*. Cambridge University Press, New York, NY, USA, 2003.
3. Thomas Bosch and Kai Eckert. Requirements on RDF Constraint Formulation and Validation. In *Proceedings of the 14th DCMI International Conference on Dublin Core and Metadata Applications (DC 2014)*, Austin, Texas, USA, 2014. http://dcevents.dublincore.org/IntConf/dc-2014/paper/view/257.
4. Thomas Bosch and Kai Eckert. Towards Description Set Profiles for RDF using SPARQL as Intermediate Language. In *Proceedings of the 14th DCMI International Conference on Dublin Core and Metadata Applications (DC 2014)*, Austin, Texas, USA, 2014. http://dcevents.dublincore.org/IntConf/dc-2014/paper/view/270.
5. Thomas Bosch, Andreas Nolle, Erman Acar, and Kai Eckert. RDF Validation Requirements - Evaluation and Logical Underpinning. *Computing Research Repository (CoRR)*, abs/1501.03933, 2015. http://arxiv.org/abs/1501.03933.
6. Richard Cyganiak, Simon Field, Arofan Gregory, Wolfgang Halb, and Jeni Tennison. Semantic Statistics: Bringing Together SDMX and SCOVO. In Christian Bizer, Tom Heath, Tim Berners-Lee, and Michael Hausenblas, editors, *Proceedings of the International World Wide Web Conference (WWW 2010), Workshop on Linked Data on the Web*, volume 628 of *CEUR Workshop Proceedings*, 2010. http://ceur-ws.org/Vol-628/ldow2010_paper03.pdf.

[55] http://rdf.insee.fr/codes/index.html
[56] http://rdf.insee.fr/codes/index.html
[57] http://rdf.insee.fr/codes/index.html

7. Thomas Hartmann, Benjamin Zapilko, Joachim Wackerow, and Kai Eckert. Constraints to Validate RDF Data Quality on Common Vocabularies in the Social, Behavioral, and Economic Sciences. *Computing Research Repository (CoRR)*, abs/1504.04479, 2015. http://arxiv.org/abs/1504.04479.
8. Mary Vardigan, Pascal Heus, and Wendy Thomas. Data Documentation Initiative: Toward a Standard for the Social Sciences. *International Journal of Digital Curation*, 3(1):107 – 113, 2008. http://www.ijdc.net/index.php/ijdc/article/view/66.